ChinaXiv合作期刊
知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum
E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

词汇链文本表示模型计算方法综述

◎ 曲云鹏[1,2,3] 王文玲[3]

[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学 北京 100049 [3] 国家图书馆 北京 100081

摘要：[目的 / 意义] 词汇链文本表示方法是一种通过词汇链对语篇中的词汇衔接关系进行建模的文本表示方法，该方法能够体现语篇中丰富的语义信息，在自动摘要、文本切分等领域得到广泛应用。[方法 / 过程] 对词汇链相关研究论文进行收集和整理，对词汇链的构建方式和消歧方法进行了归纳。词汇衔接关系的计算方法包括基于语义关联的计算方法、基于统计信息的计算方法和基于图的计算方法。词汇链构建过程中的语义消歧是很重要的过程，直接影响词汇链的构建结果和效率。[结果 / 结论] 词汇链文本表示方法结构简单、应用范围广泛。词汇链文本表示模型还存在着一些问题，如使用词典构建存在很多局限性，没有完整考虑上下文的信息等。未来词汇链模型可能会向着融合语义关系方法和统计算法、使用分布式语义加强对上下文分析等方向发展。

关键词：词汇链 词汇衔接 文本表示 自然语言处理

分类号：TP312

引用格式：曲云鹏 , 王文玲 . 词汇链文本表示模型计算方法综述 [J/OL]. 知识管理论坛 , 2016, 1(2): 136-144[引用日期]. http://www.kmf.ac.cn/paperView?id=25.

1 引言

文本表示是智能情报处理的重要环节之一，优秀的文本表示模型能充分且真实地反映文本的内容，提高智能情报处理的效果。词汇链文本表示模型是一种对语篇中的词汇衔接（lexical cohesion）关系进行建模的文本表示模型，能够体现语篇中丰富的语义信息。词汇衔接特性最早由英语语言学家 M. A. K. Halliday 和 R. Hasan 定义 [1]，指的是一段语篇中的词并不是随机组合在一起，而是围绕一个主题或事情而组织在一起。词汇衔接关系是语篇的表层特性，主要通过语篇中文本单元之间的相关性来表现，相关性包括词汇的复现现象和搭配现象。词汇的复现现象指词汇之间的语义关联，例如同义、近义、上下位、整体 - 部分关联等，词汇的搭配现象指词的共现情况，即在一定窗口距离内或某种语法规则下词汇共同出现的情况 [2]。

词汇链指的是语篇中一系列概念相关的词共同组成的词序列，词汇链文本表示模型将文本表示为几个包含有多个词的词汇链，每个词汇链中的词通过词汇衔接关系联系在一起。图 1 为一段文本中词汇链的分布情况，该文本包含两条词汇链：{sat down, rest, tired, fell asleep} 和 {beech-tree, leaf, leaves}。

作者简介：曲云鹏 (ORCID：0000-0002-1611-0904)，副研究馆员，硕士，quyp@nlc.cn；王文玲 (ORCID：0000-0002-0236-6191)，馆员，硕士。

收稿日期：2016-02-16 发表日期：2016-04-28 本文责任编辑：王善军

知识管理论坛，2016(2):136-144
DOI：10.13266/j.issn.2095-5472.2016.018

Jan sat down to rest at the foot of a huge beech-tree. Now he was so tired that he soon fell asleep; and a leaf fell on him, and then another, and then another, and before long he was covered all over with leaves, yellow, golden and brown.

图 1 文本中词汇链分布样例

词汇链能构造一个易于理解的上下文环境，有助于确定多义词在文本中的具体含义；词汇链能为文本结构以及文本一致性提供线索，有助于理解文本的大意。词汇链可以被看作是一段语篇的标志性主题词语链，这些词共同表达了同一件事情或意思，确定了词汇链就能确定一段语篇的文本结构等。词汇链文本表示模型使用广泛，不仅可有效呈现文本中的词汇衔接关系，其多种特征也可用于关键词抽取、文本切分等，例如词汇链的长度可以反映相关主题在文本中的覆盖范围，词汇链的密度可以体现语篇中相关主题的延续性，词汇链中词的分布可以体现相关主题的分布情况等。本文主要对词汇链的构建过程和构建方法进行研究和归纳，分析各种词汇链构建方法的特点并进行归类，通过对比总结出各种方法的优缺点，并探讨相关领域未来的研究方向。

2 词汇链的构建过程

在构建词汇链之前，需要先对语篇进行预处理，包括词性处理、停用词处理等，形成候选词列表。然后依照候选词出现的顺序对候选词逐一进行处理，先判断候选词 a 是否能加入已有词汇链 l。判断标准是看候选词和词汇链中的词是否有足够强的词汇衔接关系，若 a 和现有词汇链的关系满足条件，则加入；若不能加入，则新建一个词汇链并将 a 作为词汇链的第一个词。该步骤完成后会形成多个词汇链，这时根据具体需要，确定是否执行词汇链的排序、筛选、合并等操作，结果即为最终的词汇链表示模型。词汇链的构建过程如图 2 所示：

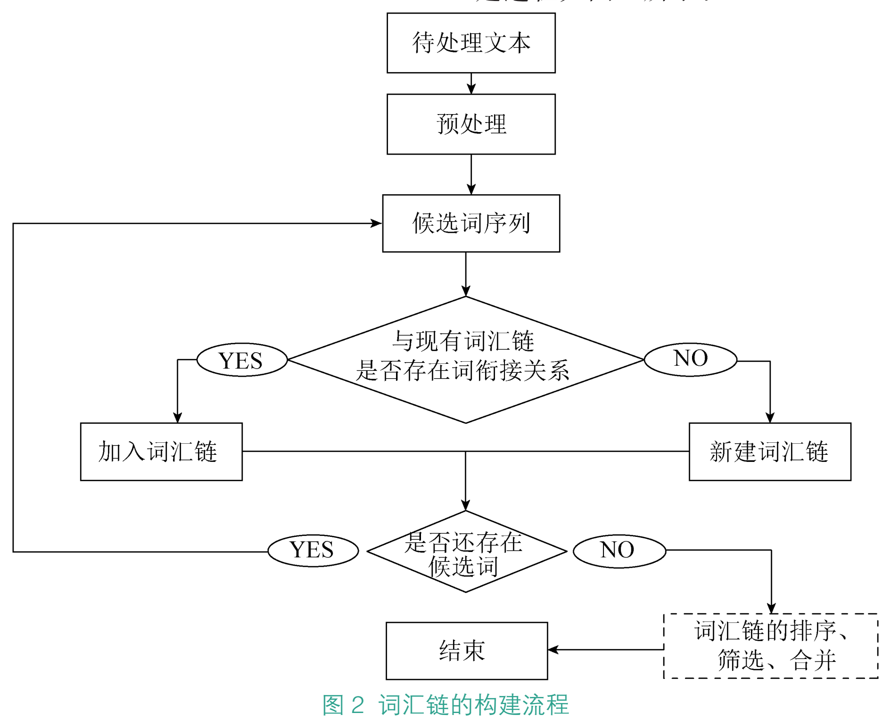


图 2 词汇链的构建流程

从词汇链的构建流程不难看出，如何寻找并计算词汇衔接关系是词汇链构建过程中的关键步骤。词汇衔接关系分为简单重复、复杂重复、简单释义、复杂释义、语义关联和非词重复 6 种类型 [3]，识别的难度从上到下按顺序逐渐增加，见表 1。词典中通常会定义一些语义关联，如上下义、同义等，可以体现一些语言单元之间的词汇衔接关系，因此早期的词汇链构建算法通常借助词典中的语义关联来对词汇衔接结构进行建模。

表 1 词汇衔接关系类型

类型	定义	样例
简单重复	词的简单重复（单复数）	bear/bears
复杂重复	有相同词根的两个词的重复，但是词性不同	historical/history quoted/quotation
简单释义	一个词可以替换另一个词，并且含义没有改变	volume/book writing/works
复杂释义	反义、两个词的关联可以推断出同第三个词的关联	hot/cold writer/writing/author teacher/teaching/instruction
语义关联	上义词，上下义，共指	bear/animal scientist/biologists
非词重复	人称代词和指示代词	he,she,it,they,this, that,these,those

另外，一词多义是自然语言最常见的特征之一，消除候选词的歧义是词汇链构建过程必不可少的步骤，语义消歧的结果体现了自然语言处理的水平。对候选词进行消歧可贯穿词汇链构建的整个过程，消除候选词歧义时机的选择将影响词汇链构建的计算复杂度和难度，同时也会影响词汇链构建的准确率。

3 词汇衔接关系计算方法

3.1 基于语义关联的计算方法

J. Morris 和 G. Hirst 首次提出词汇链算法时选择了罗杰词典（Roget's Thesaurus），他们选用了词典中的 5 种词间关系来计算词汇衔接关系：①词 a 和词 b 在词典中有相同的索引号；② a 的索引号所指向的分类有指针指向 b 的索引号所指向的分类；③ b 是 a 在词典中的标签；④ a 和 b 在同一个组；⑤ a 和 b 的索引号所指向的分类同时有指针指向另外一个分类 [4]。5 种关系的优先级按顺序递减。

WordNet 词汇数据库出现后，D. St-Onge 提出了贪婪算法，利用 WordNet 定义超强、较强和中强 3 种强度的词间关系用来计算词汇衔接关系，实现了词汇链的自动化构建 [5-6]。超强关系指某一词汇和该词在后文中的重复关系，不受距离限制。较强关系的窗口距离是 7 个句子，包括 3 种情况：①两个词属于同一个同义词集合；②两个词所属的同义词集合在 WordNet 中具有水平的层次关系；③第一个词是复合词或者短语，并且包含第二个词。中强关系的窗口距离不超过 3 个句子，在 WordNet 树形结构中的关系方向变化不超过 1 次，强度计算公式为 weight = C-（路径长度）- k*（方向变化的次数）（C 和 k 是经验常数），综合考虑了两个词在 WordNet 中关系的路径长度和语义关联的方向变化。构建词汇链时优先采用超强关系，较强关系次之，中强关系根据强度由强至弱进行采用。

WordNet 是通用的英语词典，自动化词汇链构建方法多数选择 WordNet 来进行，但是 WordNet 对英语之外的语言及特定知识领

ChinaXiv合作期刊
知识管理论坛，2016(2):136-144
DOI：10.13266/j.issn.2095-5472.2016.018

域的词汇链构建支持有限，因此很多学者开始尝试使用不同领域的专业词典或其他语言的词典来计算词汇衔接关系。在生物医学领域，L. Reeve 等人选用了美国国家医学图书馆的统一医学语言系统（Unified Medical Language System, UMLS），利用 MetaMap 工具将候选词映射为元叙词表（Metathesaurus）的概念，使用语义网络（定义了 135 种语义类型和 54 种语义关系）来计算概念语义类型之间的关联，从而构建基于叙词表概念的词汇链[7]。在中文自然语言处理领域，索红光等提出利用知网知识库来构建中文词汇链的方法[8]，通过两个词在 HowNet 中对应的两个义原所有义项的基本义原相似度、其他基本义原相似度、关系义原相似度和符号义原相似度等特征计算两个词的相似度。刘端阳等提出了基于《同义词词林》语义词典的中文词汇链构建方法[9]，首先使用候选词和词林中的释义在文本中的共现频率对多义词进行消歧，然后使用两个词在同义词词林中的分支层节点数量和分支间的距离来计算相似度。宋培彦等提出了基于概念层次网络（hierarchical network ofconcepts, HNC）的中文词汇链构建方法，利用概念层次网络中两个词义的重合度来计算两个词的语义相关度[10]。在德语自然语言处理领域，I. Cramer 等抽取 GermaNet 中的上位类关系形成上下位类树，利用两个词在上下位类树中的最短路径、绝对深度、词频等特征，用多达 8 种算法来计算词间的语义关联[11]。

以词典为工具的词汇链构建方法易于理解、便于实施，在词汇链构建过程中得到了最广泛的应用，但词典本身也有一些明显的缺点：①词典的收录范围都有一定的限制，这种限制可能是语言方面的也可能是领域范围的，还有一些新出现的词汇受到词典更新频率的限制也可能未及时被收录，词典未收录的词汇必然无法计算其语义关联；②除了体现在词典中的语义关系之外，词汇之间还有一些潜在的语义关联，如词汇的同现关系，无法通过词典来获取；③词典都是专家通过领域知识来编制的，词典质量体现了专家的专业水准，也决定了所构建词汇链的准确率。故此，单纯采用基于词典的方法构建词汇链，会遗失很多原语篇的语义信息，无法充分反映原语篇的特征。为了解决这些问题，一些研究开始融合统计语言学知识，对词汇的共现现象进行统计学分析，通过计算来发现语言单元之间的潜在语义关联，从而构建词汇链。

3.2 基于统计信息的计算方法

基于统计信息的词汇链构建方法主要有两类：第一类是对语料进行统计语言学分析形成知识库，然后利用知识库计算对象文本的相似度来判断词汇衔接关系。第二类是直接利用词汇共现关系，经过一定的变换来计算语言单元之间的相似度，作为识别词汇衔接关系的基础。

第一类方法主要是通过大规模的语料分析，对语料中词汇共现的情况进行统计、分析和计算，形成共现关系知识库，从而进行词汇链构建。G. Dias 等提出一种与语言无关的基于动态知识库的方法来计算候选词相似度，先计算语料中文本单元的上下文相似度矩阵，并将其作为初始参数输入基于极的重叠聚类算法（pole-based overlapping clustering algorithm）对矩阵进行聚类，形成一个可以揭示类别之间语义关联的知识库，然后利用知识库中语言单元之间的关联来计算词汇衔接关系[12]。M. Marathe 等使用概念距离的分布式测量（distributional measures of concept distance）方法来计算候选词与现有词汇链以及词汇链之间的语义距离，作为词汇衔接关系的表现[13]。概念距离的分布式测量是一种融合了词典语义关系和词汇共现的计算方法，使用该方法进行自动文本切分，能获得比较优化的结果。

第二类方法直接使用目标语篇中的词汇共现情况来计算候选词之间的词汇搭配关系。S. Remus 等使用了 3 种 LDA（latentDirichlet allocation）概率主题模型来计算候选词的语义相似度，将概率分布于同一个主题的词归于同一个词汇链[14]。叶春蕾等使用领域关键词作为

词汇链的初始词，通过计算候选词与领域关键词的 E 指数来判断是否将候选词加入词汇链。E 指数是一种基于同段共现分析的关联度计算指标，用来分析词语之间的语义关联强度 [15]。

基于语义关联的计算和基于统计信息的计算从两个角度来计算词汇衔接关系，无法直接比较其效果的优劣。从计算复杂度来说，基于词典的语义关联的计算方法复杂度较低，准确性较高，但是受到词典本身缺点的影响，可能丢失一些语义特征。基于大规模统计分析的计算方法能探测到两个术语之间的潜在语义关联，可以弥补基于语义关联的构建方法在该方面存在的不足。基于词汇共现的计算方法可以探测到单篇文档中的术语之间的特殊语义关联，这对基于大规模统计分析的计算方法是一个补充。基于统计信息的计算方法需要大量的计算，同时知识库的构建需要大量语料的支持，计算复杂度要远高于基于语义关联的方法。

3.3 基于图的计算方法

最早的词汇链构建方法都是依照候选词出现的顺序来构建词汇链，例如 J. Morris 和 G. Hirst 的词汇链构建算法 [4]、贪婪算法 [16] 和 R. Barzilay 等提出的非贪婪算法 [17] 等。顺序构建词汇链的问题在于，处理候选词时只能计算该词与已处理的候选词之间的关系，无法计算该词与其后出现的候选词之间的关系，然而相同的候选词集合如果以不同顺序进行处理，可能会得到不同的结果，从而影响词汇链构建的准确性。基于图的构建方式是将所有的候选词取出，分别计算每对候选词之间的语义关联，形成图结构，然后使用图聚类等算法，对图中的边进行消减，从而形成最终的词汇链。

O.MEDELYAN 提出用图聚类的方法来构造词汇链 [18]，先顺序处理每个候选词，如果候选词可以加入多个词汇链，则将这些词汇链合并。将所有词汇链转化为图后，将图中最长的任意两点间的最小路径距离定义为图标度(graph diameter)，对于图标度大于 3 的弱链，利用图聚类的方法来识别弱链中的高凝聚子图作为计算结果，图标度小于 3 的链可以直接作为计算结果。S. KATIYAR 等也提出了一种基于图的词汇链构建方法 [19]，首先将候选词作为顶点构建一个图，图的边及权重通过候选词之间的关系强度来计算，候选词之间的关系强度则由它们在 WordNet 分类法中的距离决定。图构建完成后，对每个顶点 v_i 将图中所有和 v_i 相连的顶点加入列表 W_i，对 W_i 中的每个候选词 w_j，计算 w_j 同 W_i 中其他候选词的语义关联强度，并将这些强度相加作为三元组 $<v_i, w_j, Score>$ 中的权重。计算完成后，各顶点与图中关联关系最多的顶点之间的关系会更强，确保各顶点的准确语义得到更多的体现。将计算完的图分解为不相交的最长子链，即不重复的子图，则得到最终的词汇链。

基于图的词汇链构建过程不考虑候选词的顺序关系，将所有候选词之间的相互关系映射为加权图，再利用图聚类等算法筛选满足设定条件的词汇关系，形成最终的词汇链文本表示模型。基于图的词汇链构建能发现顺序构建方法可能丢失的语义，也尽可能通过上下文语义去消减可能造成的歧义，但是仍然可能会造成对文本的错误表示。

4 词汇链构建中的语义消歧

多义词是自然语言中常见的现象，在词汇链构建过程中，如果可以为多义词选择准确的含义消除歧义，必然会提高词汇链构建的准确率，降低词汇链构建的复杂度。语义消歧根据其在词汇链构建过程中的时机，可以分为提前消歧、构建时消歧和构建后消歧。

4.1 提前消歧

提前消歧指的是在构建词汇链之前，对候选词进行消歧，确定候选词含义。CHAD 算法 [20] 采用了提前消歧的策略，主要思路是在有序相邻的 3 个词中，如果第一、二个词的含义已确定，则第三个词的含义可以通过计算来确定。利用 CHAD 算法可以侦测到文本中词义连续的停滞状态，即若相邻的两个词的含义完全没有

知识管理论坛，2016(2):136-144
DOI：10.13266/j.issn.2095-5472.2016.018

相交，此时词汇衔接现象就会停止。F.Y. YE 等人的提前消歧算法综合考虑了窗口长度和关系强度，针对候选词的每种词义，计算长度为 6 的窗口距离中所有候选词与该词的关联强度总和，选择关联强度总和最大的词义作为该候选词的词义，从而达到消歧的目的[21]。

4.2 构建时消歧

构建时消歧指构建词汇链时，同时判断候选词词义和候选词的词汇链归属，是词汇链计算中较为常用的消歧方法。在判断词汇链归属时，计算候选词所有词义同所有词汇链的关系，从中选择满足条件的词义作为候选词的词义并加入相应的词汇链，选择词义时可以考虑同某个词汇链关联最多的词义，也可以考虑加权语义关联最大的词义。非贪婪算法采用的是构建时消歧，先从 WordNet 词汇数据库中抽取候选词的多个词义，选择与词汇链中其他成员间的语义关系数量最多的词义作为该候选词的词义，并计算该词义与词汇链中成员的语义关联的权值，作为词汇衔接关系的权值，用于判断是否将候选词加入词汇链[17]。

4.3 随后消歧

随后消歧指的是，在构建词汇链的过程中保留所有可能语义形成的词汇链作为候选链，在词汇链构建完之后，依据某种条件对所有的链实施二次处理进行消歧。元链（metachain）算法采用了随后消歧的方法[22-24]，先将 WordNet 词汇数据库进行扁平化处理，抽取所有的词义作为每个元链的开头词义，对 WordNet 中的语义关系赋予权值并进行词汇链构建，构建完的元链结果列出了文本所有可能的语义解释。随后对元链进行筛选，只选择每个候选词所有词义中对所在词汇链的语义权重贡献最大的词汇链进行保留，同时从其他词汇链中删除该词，剩下的词汇链作为最终结果。M.Galley 等延续了基于元链的词汇链构建思路，使用“一词一义”（one sense per course）的方法进行消歧，即假设每个词在一个语篇中只有一个含义，当含义确定时，词汇链中所有的同一个词的实例都使用同一个含义[25]。

提前消歧是在词汇链构建之前即对候选词赋予明确的词义，这大大降低了词汇链构建的计算复杂度，但是词汇链构建的效果完全取决于消歧所使用的算法，如果消歧所使用的算法效果不甚理想，会导致候选词词义错误，形成的词汇链质量不高。构建时消歧需要在构建词汇链的同时考虑候选词的上下文确定候选词的词义，计算复杂度较高，但是由于可以动态联系上下文进行消歧，消歧的准确性较好。随后消歧相对于构建时消歧，可以更全面考虑候选词的上下文情况，但是需要计算所有候选词的每种词义所产生的效果，因此其计算复杂度是最高的。

5 相关研究评述

以上对词汇链构建过程中的词衔接关系计算和消歧方法进行了归纳。可以发现，词汇链可以有效地识别文本中词衔接关联的延续性。目前词汇链构建方法中还有一些不足，需要进一步研究解决。

（1）基于词典的方法有一些不足，应尝试融合基于统计信息和基于词典的方法进行词汇链的构建。通过词典中的语义关系来构建词汇链有很多不足：第一，基于词典的方法对于词典的依赖性很强，词典的语种、质量、更新频率等都会影响词汇链的构建。第二，词典无法识别候选词之间的潜在关联。语言的发展很迅速，词典由于需要人工来维护，始终落后于语言发展，一些新的关联可能无法及时收录。第三，如果词典中无法查到某些候选词，那么在构建词汇链时，只好抛弃这些候选词。第四，词汇衔接关系包括基于语义关联的词复现关系和基于共现的词汇搭配关系，单一使用基于语义关联或者基于统计信息的算法都可能造成衔接关系识别的不完整。基于统计的方法可以识别一些候选词之间的潜在关联，也可以探测候选词之间的搭配关系。可见，基于词典的方法和基于统计信息的方法可以相互补充。尝试融合

基于词典的方法和基于统计信息的方法，尽可能完整地计算词衔接关系，应是词汇链构建的主要研究方向之一。

（2）词汇链构建过程中，对候选词上下文的考虑不够，可以尝试使用分布式语义（distributional semantics）模型等方法，充分考虑上下文对候选词含义的影响。现有的词汇链构建算法多采用名词作为候选词，但是动词、形容词等其他词性的词也存在词汇衔接关系，对语篇的语义有一定的指示作用，对周围的候选词的含义也有影响，忽略了这部分关联可能会造成文本表示模型对原语篇的误解。一些研究者开始尝试采用动词、形容词和副词作为候选词[26-28]，但是并没有将这些因素作为上下文环境来考虑。分布式语义模型的基本思想是通过大量语料训练，掌握术语与上下文中其他词汇的共现关系，并以此作为术语常用上下文信息，通过对比术语的上下文信息，来判断术语之间的关联强度[29]。因此，可以尝试研究候选词和上下文中的形容词、副词等词性的词的分布式语义关系，用于构建词汇链，提升候选词的消歧效果，增强对词典中没有收录的词的识别效果。

（3）对于开头词的选择还需要进行深入研究。在顺序构建词汇链的方法中，词汇链开头词的选取会对词汇链的构建产生深远的影响，如果语篇的第一个词和语篇所要表达的主题关系不大甚至无关，在顺序处理候选词时，可能会造成候选词的消歧错误，导致词汇链构建不准确。元链算法中尝试将 WordNet 展开作为词汇链的开头词，但是算法过于复杂。在词汇链构建中，可以考虑通过统计信息选择词频较高的非停用词作为词汇链的开头词，在提高词汇链构建效果的基础上，减少计算复杂度。

5 结语

本文通过对现有相关文献进行调研梳理，将词汇链构建方法和计算过程中的消歧方法进行分类，分别对其优缺点进行分析和比较，阐述了词汇链构建在未来的研究及应用中的发展方向。词汇链文本表示方法结构简单，应用范围广泛，除了文本切分、自动摘要等领域外，词汇链还可以应用于文本过滤[30]、自动问答[31]、拼写错误识别[5]和情感识别[32]等领域，是值得深入研究的文本表示方法。随着研究进一步加深，词汇链文本表示模型将会得到更广泛的应用。

参考文献：

[1] HALLIDAY M A K, HASAN R. Cohesion in English (English Language)[M]. London:Longman, 1976.

[2] HE Q. A study of lexical cohesion theory in reading comprehension[J]. International journal of English linguistics, 2014, 4(6): 143-150.

[3] TOWNS S G, TODD R W. Disunity in cohesion: how purpose affects methods and results when analyzing lexical cohesion[C]//AROONMANAKUNW, BOONKWAN P, SUPNITHI T. Proceedings of the28th Pacific Asia conference on language, information and computation(PACLIC 28).Bangkok: Department of Linguistics, Faculty of Arts, Chulalongkorn University, 2014: 513-521.

[4] MORRIS J, HIRST G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text[J]. Computational linguistics, 1991, 17(1): 21-48.

[5] ST-ONGE D. Detecting and correcting malapropisms with lexical chains[D]. Toronto: University of Toronto, 1995.

[6] HIRST G, ST-ONGE D. Lexical chains as representations of context for the detection and correction of malapropisms[C]// FELLBAUM C. WordNet: an electronic lexical database. Cambridge: The MIT Press, 1995: 305-331.

[7] REEVE L, HAN H, BROOKS A D. BioChain: lexical chaining methods for biomedical text summarization[C]// ACM Special Interest Group on Applied Computing. Proceedings of the 2006 ACM symposium on applied computing. New York: ACM, 2006: 180-184.

[8] 索红光，刘玉树，曹淑英 . 一种基于词汇链的关键词抽取方法 [J]. 中文信息学报，2006, 20(6): 25-30.

[9] 刘端阳，王良芳 . 基于语义词典和词汇链的关键词提取算法 [J]. 浙江工业大学学报，2013, 41(5): 545-551.

[10] 宋培彦，杨代庆 . 基于语义网络的中文词汇链构造方法 [J]. 图书情报工作，2011, 55(22): 26-29.

[11] CRAMER I, FINTHAMMER M. An evaluation procedure for word net based lexical chaining:

methods and issues[EB/OL]. [2016-03-15].http://www.irene-cramer.de/resources/CramerFinthammer_cameraVersion_2007-10-31.pdf.

[12] DIAS G, SANTOS C, CLEUZIOU G. Automatic knowledge representation using a graph-based algorithm for language-independent lexical chaining[C]//Proceedings of the workshop on information extraction beyond the document.Stroudsburg: Association for Computational Linguistics, 2006: 36-47.

[13] MARATHE M, HIRST G. Lexical chains using distributional measures of concept distance[C]//GELBUKH A. Computational linguistics and intelligent text processing. Heidelberg: Springer Berlin Heidelberg, 2010: 291-302.

[14] REMUS S, BIEMANN C. Three knowledge-free methods for automatic lexical chain extraction[C]//Proceedings of NAACL-HLT 2013. Stroudsburg: Association for Computational Linguistics, 2013: 989-999.

[15] 叶春蕾，冷伏海．基于词汇链的路线图关键词抽取方法研究 [J]. 现代图书情报技术，2013(1): 50-56.

[16] HIRST G, ST-ONGE D. Lexical chains as representations of context for the detection and correction of malapropisms[C]//FELLBAUM C. WordNet: an electronic lexical database. Cambridge: MIT press, 1995: 305-331.

[17] BARZILAY R, ELHADAD M. Using lexical chains for text summarization[C]//MANI I,MAYBURY M T.Advances in automatic text summarization. Cambridge: MIT Press, 1999: 357-380.

[18] MEDELYANO. Computing lexical chains with graph clustering[C]//BIEMANN C, SERETANV.Proceedings of the 45th annual meeting of the ACL: student research workshop. Stroudsburg: Association for Computational Linguistics, 2007: 85-90.

[19] KATIYAR S, BORGOHAIN S K.A novel approach towards automatic text summarization using lexical chains[J]. International journal on recent and innovation trends in computing and communication, 2015, 3(8): 5115-5121.

[20] TATAR D, MIHIS A D, CZIBULA G S. Lexical chains segmentation in summarization[C]//Proceedings of the 2008 10th international symposium on symbolic and numeric algorithms for scientific computing. Washington: IEEE Computer Society, 2008: 95-101.

[21] YE FY, CHEN Y, LUO X, et al. Research on topic segmentation of Chinese text based on lexical chain[C]//Proceedings of 2012 IEEE 12th international conference on computer and information technology. Washington: IEEE Computer Society, 2012: 1131-1136.

[22] SILBER H G, MCCOY K F. Efficiently computed lexical chains as an intermediate representation for automatic text summarization[J]. Computational linguistics, 2002, 28(4): 487-496.

[23] SILBER H G, MCCOY K F. An efficient text summarizer using lexical chains[C]//Proceedings of the first international conference on natural language generation - volume 14. Stroudsburg: Association for Computational Linguistics, 2000: 268-271.

[24] SILBER H G, MCCOY K F. Efficient text summarization using lexical chains[C]//Proceedings of the 5th international conference on intelligent user interfaces.New York: ACM, 2000: 252-255.

[25] GALLEY M, McKeown K, Improving word sense disambiguation in lexical chaining[C]// Proceedings of the 18th international joint conference on artificial intelligence. San Francisco: Morgan Kaufmann Publishers Inc.,2003: 1486-1488.

[26] SANGEETHA S, THAKUR R S, AROCK M. Event detection using lexical chain[C]//LOFTSSONH. Advances in natural language processing. Berlin: Springer Berlin Heidelberg, 2010: 314-319.

[27] NOVISCHI A, MOLDOVAN D. Question answering with lexical chains propagating verb arguments[C]//Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the ACL.Stroudsburg: Association for Computational Linguistics, 2006: 897-904.

[28] CHEN J, LIU J, YU W, et al. Combining lexical stability and improved lexical chain for unsupervised word sense disambiguation[C]//Proceedings of the 2009 second international symposium on knowledge acquisition and modeling - volume 01. Washington: IEEEComputer Society, 2009: 430-433.

[29] PADÓ S, LAPATA M. Dependency-based construction of semantic space models[J]. Computational linguistics, 2007, 33(2): 161-199.

[30] LI S, YOU W, LI T, et al. Lexical-chain and it's application in text filtering[C]//Proceedings of the international conference on information technology: coding and computing.Washington: IEEE Computer Society, 2004: 288-292.

[31] MOLDOVAN D, NOVISCHI A. Lexical chains for question answering[C]//Proceedings of the 19th international conference on computational linguistics - volume 1. Stroudsburg:Association for Computational

Linguistics, 2002:1-7.
[32] KUMAR M N, SURESH R. Emotion detection using lexical chains[J]. International journal of computer applications, 2012, 57(4): 1-4.

作者贡献说明：
曲云鹏：提出研究思路，负责论文起草。
王文玲：负责论文的撰写和最终版本的修订。

An Overview on the Computing Method of the Lexical Chain Text Representation Model

Qu Yunpeng[1,2,3] Wang Wenling[3]
[1]National Science Library, Chinese Academy of Sciences, Beijing 100190
[2]University of Chinese Academy of Sciences, Beijing 100049
[3]National Library of China, Beijing 100081

Abstract: [Purpose/significance] Text representation is an important step in intelligence processing. An excellent text representation model can reflect the document content precisely and sufficiently. Besides, it can improve the processing effect. It can be broadly applied in the fields of automatic abstracting and text segmentation. **[Method/process]** In this article, we collected the related documents and analyzed them. The construction methods and disambiguation in the lexical chain computing were classified and concluded. The computing method of the lexical chain relation included the computing method based on semantic association, the computing method based on statistical information and the computing method based on charts. The semantic disambiguation was important in the construction of the lexical chain, which directly affected the results and efficiency of the lexical chain construction. **[Result/conclusion]** The lexical chain text representation can be easily constructed and broadly applied. There are still some problems in the text representation model of the lexical chain. For example, there are many limitations to construct it by dictionaries, which does not take the context into consideration. The lexical chain model will possibly develop towards the fusion semantic relation method, the statistical algorithm and the context analysis of distributed semantics in the future.

Keywords: lexical chains lexical cohesion text representation natural language processing